



# Towards a Rigorous Assessment of Systems Biology Models: The DREAM3 Challenges

## Citation

Prill, Robert J., Daniel Marbach, Julio Saez-Rodriguez, Peter K. Sorger, Leonidas G. Alexopoulos, Xiaowei Xue, Neil D. Clarke, Gregoire Altan-Bonnet, and Gustavo Stolovitzky. 2010. Towards a Rigorous Assessment of Systems Biology Models: The DREAM3 Challenges. PLoS ONE 5(2)# e9202.

## Published Version

doi:10.1371/journal.pone.0009202

## Permanent link

<http://nrs.harvard.edu/urn-3:HUL.InstRepos:7628414>

## Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA>

## Share Your Story

The Harvard community has made this article openly available.  
Please share how this access benefits you. [Submit a story](#).

[Accessibility](#)

## Overview

# Towards a Rigorous Assessment of Systems Biology Models: The DREAM3 Challenges

Robert J. Prill<sup>1</sup>, Daniel Marbach<sup>2</sup>, Julio Saez-Rodriguez<sup>3,4</sup>, Peter K. Sorger<sup>3,4</sup>, Leonidas G. Alexopoulos<sup>5</sup>, Xiaowei Xue<sup>6</sup>, Neil D. Clarke<sup>6</sup>, Gregoire Altan-Bonnet<sup>7</sup>, Gustavo Stolovitzky<sup>1\*</sup>

**1** IBM T. J. Watson Research Center, Yorktown Heights, New York, United States of America, **2** Laboratory of Intelligent Systems, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, **3** Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, United States of America, **4** Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **5** Department of Mechanical Engineering, National Technical University of Athens, Athens, Greece, **6** Genome Institute of Singapore, Singapore, Singapore, **7** Program in Computational Biology and Immunology, Memorial Sloan-Kettering Cancer Center, New York, New York, United States of America

**Background:** Systems biology has embraced computational modeling in response to the quantitative nature and increasing scale of contemporary data sets. The onslaught of data is accelerating as molecular profiling technology evolves. The Dialogue for Reverse Engineering Assessments and Methods (DREAM) is a community effort to catalyze discussion about the design, application, and assessment of systems biology models through annual reverse-engineering challenges.

**Methodology and Principal Findings:** We describe our assessments of the four challenges associated with the third DREAM conference which came to be known as the DREAM3 challenges: signaling cascade identification, signaling response prediction, gene expression prediction, and the DREAM3 *in silico* network challenge. The challenges, based on anonymized data sets, tested participants in network inference and prediction of measurements. Forty teams submitted 413 predicted networks and measurement test sets. Overall, a handful of best-performer teams were identified, while a majority of teams made predictions that were equivalent to random. Counterintuitively, combining the predictions of multiple teams (including the weaker teams) can in some cases improve predictive power beyond that of any single method.

**Conclusions:** DREAM provides valuable feedback to practitioners of systems biology modeling. Lessons learned from the predictions of the community provide much-needed context for interpreting claims of efficacy of algorithms described in the scientific literature.

## Introduction

Computational models of intracellular networks are a mainstay of systems biology. Researchers have used a variety of algorithms to deduce the structure of very different biological and artificial networks [1] and have evaluated their success using various metrics [2–8]. What is needed is a fair comparison of the strengths and weaknesses of the methods and a clear sense of the reliability of the network models they produce.

The Dialogue on Reverse Engineering Assessment and Methods (DREAM) project “takes the pulse” of the current state of the art in systems biology modeling [9,10]. DREAM is organized around annual reverse-engineering challenges whereby teams download data sets from recent unpublished research, then attempt to recapitulate some withheld details of the data set. A challenge typically entails inferring the connectivity of the molecular networks underlying the measurements, predicting withheld measurements, or related reverse-engineering tasks. Assessments of the predictions are blind to the methods and identities of the participants.

The format of DREAM was inspired by the Critical Assessment of techniques for protein Structure Prediction (CASP) [11]

whereby teams attempt to infer the three-dimensional structure of a protein that has recently been determined by X-ray crystallography but temporarily withheld from publication for the purpose of creating a challenge. Instead of protein structure prediction, DREAM is focused on network inference and related topics that are central to systems biology research. While no single test of an algorithm is a panacea for determining efficacy, we assert that the DREAM project fills a deep void in the validation of systems biology algorithms and models. The assessment provides valuable feedback for algorithm designers who can be lulled into a false sense of security based on their own internal benchmarks. Ultimately, DREAM and similar initiatives may demystify this important but opaque area of systems biology research so that the greater biological research community can have confidence in this work and build new experimental lines of research upon the inferences of algorithms.

## Evolution of the DREAM Challenges

At the conclusion of the second DREAM conference [10], a few voices of reason suggested that reverse-engineering challenges should not be solely focused on the network inference. As the argument goes, only that which can be measured should be

**Citation:** Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, et al. (2010) Towards a Rigorous Assessment of Systems Biology Models: The DREAM3 Challenges. PLoS ONE 5(2): e9202. doi:10.1371/journal.pone.0009202

**Editor:** Mark Isalan, Center for Genomic Regulation, Spain

**Received:** November 19, 2009; **Accepted:** January 19, 2010; **Published:** February 23, 2010

**Copyright:** © 2010 Prill et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** G.S. and R.P. acknowledge support of the National Institutes of Health (NIH) Roadmap Initiative, the Columbia University Center for Multiscale Analysis Genomic and Cellular Networks (MAGNet), and the IBM Computational Biology Center. L.A. acknowledges funding from the Marie Curie MIRG-CT-2007-46531 grant. J.R. acknowledges funding from NIH P50-GM68762 and U54-CA11296 grants. G.A.-B. acknowledges funding from a National Science Foundation (NSF) grant 0848030 and NIH grant AI083408. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: gustavo@us.ibm.com

predicted. Since knowledge of biological networks is actually a model in its own right, it may be counterproductive to evaluate networks for which no ground truth is known. We agree that the positivist viewpoint has merit both as a matter of philosophy and practicality. Some of the DREAM3 challenges reflect this attitude, which was a shift from previous challenges which were squarely focused on network inference.

Nevertheless, the systems biology community continues to assert—through funding opportunities, conference attendance, and the volume of publications—that network inference is a worthwhile scientific endeavor. Therefore, DREAM continues to provide a venue for vetting algorithms that are claimed to reverse-engineer networks from measurements. Despite the above mentioned criticisms, network inference challenges are a mainstay of DREAM. To contend with the criticism that no ground truth is known for molecular networks, the organizers must occasionally tradeoff realism for truth—generating *in silico* (i.e., simulated) data is one way that this problem is mitigated.

We describe the results of the DREAM3 challenges: signaling cascade identification, signaling response prediction, gene expression prediction, and *in silico* network inference. The fourth challenge was similar to the DREAM2 *in silico* network inference challenge [10] which enabled a cursory analysis of progress (or lack thereof) in the state of the art of network inference. The best-performer strategies in each challenge are described in detail in accompanying publications in this PLoS ONE Collection. Here, our focus is the characterization of the efficacy of the reverse-engineering community as a whole. The results are mixed: a handful of best-performer teams were identified, yet the performance of most teams was not all that different from random.

In the remainder of the Introduction we describe each of the four DREAM3 challenges. In Results and Discussion, we summarize the results of the prediction efforts of the community, identify best-performer teams, and analyze the impact of the community as a whole. Readers interested in a particular challenge can read the corresponding sub-sections without loss of continuity. Finally, we present the conclusions of this community-wide experiment in systems biology modeling.

## The DREAM3 Challenges

In this section we describe the challenges as they were presented to the partici-

pants. Additionally, we elaborate on the experimental methods at a level of detail that was not provided to the participants. We then go on to describe the basis of the assessments—“gold standard” test sets, scoring metrics, null models and *p*-values. The complete challenge descriptions and data are archived on the DREAM website [12].

### Signaling Cascade Identification

The signaling cascade identification challenge explored the extent to which signaling proteins are identifiable from flow cytometry data. Gregoire Altan-Bonnet of Memorial Sloan-Kettering Cancer Center generously donated the data set consisting of pairwise measurements of proteins that compose a signaling pathway in T cells [13]. The data producer, cell type, and protein identities were not disclosed to participants until the results of the challenge were released.

Protein concentrations in a signaling network were measured in single cells by antibody staining and flow cytometry. Participants were provided a network diagram (Figure 1) and pairwise measurements of four signaling proteins (denoted x1, x2, x3, x4) obtained from single cells. The pairs of proteins (x1, x4), (x2, x4), and (x3, x4) were simultaneously measured in separate assays. The task was to identify each of the measured proteins (x1, x2, x3, x4) from among the seven molecular species (complex, phosphorylated complex, protein, phosphorylated protein, kinase, phosphatase, and activated phosphatase).

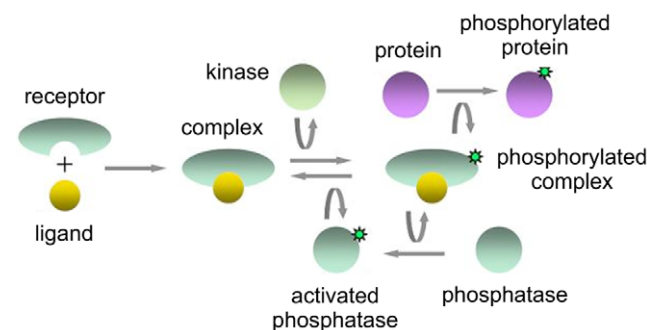
The experimental setup allowed for external control over the signaling network through the ligand that binds to the membrane-bound receptor. Two types of ligands, weak and strong (i.e., with different potency), in different concentrations,

were used. Five concentrations of strong ligand (including none) and seven concentrations of weak ligand (including none) were applied to approximately  $10^4$  cells in separate experiments. In total, data from 36 experiments corresponding to the various combinations of quantified proteins, ligand type, and ligand concentration were provided. The biological motivation of the T cell experiment is discussed in [13].

**Basis of assessment.** Participants were instructed to identify each of each of the four measured proteins (x1, x2, x3, x4) as a molecular species (kinase, phosphatase, etc.). Each measurement could only be identified as a single molecular species, and each molecular species could be assigned to at most one measurement. For example, if measurement x1 was identified as the kinase then no other measurement could also be identified as the kinase. Submissions were scored by the probability that a random assignment table would result in as many correct identifications as achieved by the participant.

There are 840 possible assignment tables for seven molecular species and four measurements (i.e.,  $7 \times 6 \times 5 \times 4$ ). The probability of guessing the gold standard assignment table by chance is  $1/840$ , which we denote  $\alpha$ . By enumerating the 840 tables and counting the number of correct (or incorrect) assignments in each table, we obtain the probability of correctly identifying four, three, two, one, or zero molecular species. It can be shown that the probability  $P(\cdot)$  of making some number of correct identifications is exactly

$$P(4 \text{ correct}) = \frac{1}{7 \times 6 \times 5 \times 4} = \alpha$$



**Figure 1. The objective of the signaling cascade identification challenge was to identify some of the molecular species in this diagram from single-cell flow cytometry measurements.** The upstream binding of a ligand to a receptor and the downstream phosphorylation of a protein are illustrated.  
doi:10.1371/journal.pone.0009202.g001

$$P(3 \text{ correct}) = \binom{4}{1} \frac{1}{7 \times 6 \times 5} \times \frac{3}{4} = 12\alpha$$

$$P(2 \text{ correct}) = \binom{4}{2} \frac{1}{7 \times 6} \left( \frac{1}{5} + \frac{3}{5} \times \frac{3}{4} \right) = 78\alpha$$

$$P(1 \text{ correct}) = \binom{4}{3} \frac{1}{7} \left[ \frac{2}{6} \left( \frac{1}{5} + \frac{4}{5} \times \frac{3}{4} \right) + \frac{3}{6} \left( \frac{1}{5} + \frac{3}{5} \times \frac{3}{4} \right) \right] = 284\alpha$$

$$P(0 \text{ correct}) = \frac{3}{7} \left[ \frac{2}{6} \left( \frac{1}{5} + \frac{4}{5} \times \frac{3}{4} \right) + \frac{4}{6} \left( \frac{1}{5} + \frac{3}{5} \times \frac{3}{4} \right) + \frac{3}{7} \left[ \frac{2}{6} \left( \frac{1}{5} + \frac{4}{5} \times \frac{3}{4} \right) + \frac{3}{6} \left( \frac{1}{5} + \frac{3}{5} \times \frac{3}{4} \right) \right] \right] = 465\alpha$$

In addition to assigning a score to each team, we characterized the efficacy of the community as a whole. For example, what is the probability that five teams would correctly identify the same protein? To compute *p*-values for community-wide outcomes such as this we used the binomial distribution which is explained in Results.

### Signaling Response Prediction

The signaling response prediction challenge explored the extent to which the responses to perturbations of a signaling pathway can be predicted from a set of training data consisting of perturbations (environmental cues and signaling protein inhibitors) and their responses. Peter Sorger of Harvard Medical School generously donated the data for this challenge consisting of time-series measurements of a signaling network measured in human hepatocytes [14,15]. The task entailed predicting some phosphoprotein and cytokine measurements that were withheld from the participants.

Approximately 10,000 fluorescence measurements proportional to the concentration of intracellular phosphorylated proteins and extracellular cytokines were acquired in normal human hepatocytes and the hepatocellular carcinoma cell line HepG2 using the Luminex (Austin, TX) 200

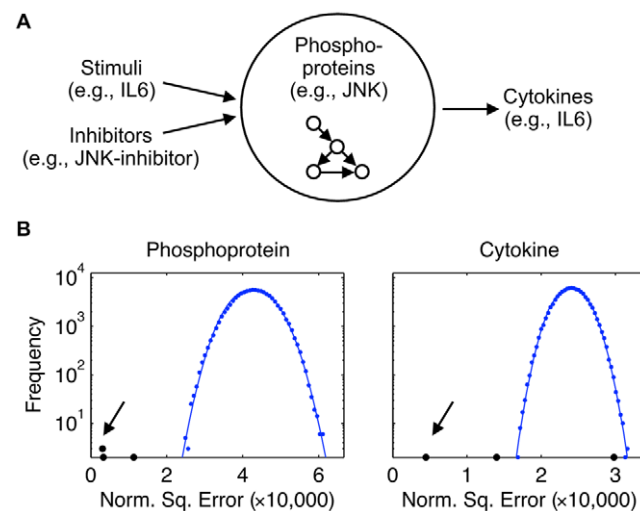
xMAP system. The data set consisted of measurements of 17 phosphoproteins at 0 minutes, 30 minutes, and 3 hrs following stimulation/perturbation of the two cell types. Additionally, 20 cytokines were quantified at 0 minutes, 3 hours, and 24 hours following stimulation/perturbation of the two cell types. Data were processed and visualized using the open-access MATLAB-based software, DataRail [16]. The cell types and protein identities were disclosed so that participants could draw upon the existing signal transduction literature.

In each experiment, a combination of a single chemical stimulus and a single chemical perturbation to the signaling network were simultaneously applied, and measurements of either the signaling network proteins or cytokines were taken (Figure 2A). Seven stimuli were investigated: INF $\gamma$ , TNF $\alpha$ , IL1 $\alpha$ , IL6, IGF-I, TGF $\alpha$ , and LPS (Table 1). Also, seven chemical inhibitors of specific signaling proteins were investigated, which selectively inhibited the activities of MEK12, p38, PI3K, IKK, mTOR, GSK3, or JNK. All pairs of stimulus/inhibitor combinations (a total of 64) were applied to both cell types and measurements of fluorescence for individual proteins or cytokines were taken at the indicated time points. Fluorescence was reported in arbitrary units from 0 to ~29000. The upper limit corresponded to saturation of the detector. Signal

intensity below 300 was considered noise. Fluorescence intensity was approximately linear with concentration in the mid-dynamic range of the detector.

The challenge was organized in two parts that were evaluated separately: the phosphoprotein subchallenge and the cytokine subchallenge. The complete data set (training and test) in the signaling response prediction challenge was composed of fluorescence measurements of phosphoproteins and cytokines in cells exposed to pairwise combinations of eight stimuli and eight signaling-network-protein inhibitors, for a total of 64 stimulus/inhibitor combinations (including zero concentrations). Fifty-seven of the combinations composed the training set, and seven combinations composed the test set. The phosphoprotein subchallenge solicited predictions for 17 phosphoproteins, in two cell types (normal, carcinoma), at two time points, under seven combinatoric stimulus/inhibitor perturbations for a total of 476 predictions. Likewise, the cytokine subchallenge solicited predictions for 20 cytokines for a total of 560 predictions. The biological motivation for the hepatocyte experiment is described in [14,15].

**Basis of assessment.** Assessment of the predicted measurements was based on a single metric, the normalized squared error over the set of predictions in each subchallenge,



**Figure 2. The objective of the signaling response prediction challenge was to predict the concentrations of phosphoproteins and cytokines in response to combinatorial perturbations the environmental cues (stimuli) and perturbations of the signaling network (inhibitors).** (a) A compendium of phosphoprotein and cytokine measurements was provided as a training set. (b) Histograms (log scale) of the scoring metric (normalized squared error) for 100,000 random predictions were approximately Gaussian (fitted blue points). Significance of the predictions of the teams (black points) was assessed with respect to the empirical probability densities embodied by these histograms. Scores of the best-performer teams are denoted with arrows.  
doi:10.1371/journal.pone.0009202.g002

**Table 1.** The signaling response prediction challenge solicited predictions of the concentrations of 17 phosphoproteins and 20 cytokines.

Phosphoproteins (17)	Inhibitors (7)	Cytokines (20)	Stimuli (7)
Akt		IL1b	
ERK1/2		IL4	
GSK-3alpha/beta		IL6	IL6
IkappaB-alpha		IL8	
JNK	JNK-i	IL10	
p38 MAPK	p38-i	IL15	
p70 S6 kinase		GCSF	
p90RSK		GMCSF	
STAT3		IP10	
c-Jun		MCP1	
CREB		MIP1a	
Histone H3		MIP1b	
HSP27		PDGFbb	
IRS-1		RANTES	
MEK1		VEGF	
p53		GROa	
STAT6		ICAM1	
	MEK12-i	MIF	
	PI3K-i	MIG	
	IKK-i	SDF1a	
	mTOR-i		INFg
	GSK3-i		TNFa
			IL1a
			IGF-I
			TGFa
			LPS

The data set underlying this challenge consisted of phosphoprotein and cytokine concentrations in response to 49 combinatoric perturbations of seven protein-specific inhibitors and seven stimuli.  
doi:10.1371/journal.pone.0009202.t001

$$\text{Norm. Sq. Error} = \sum_{i=1}^n \frac{(\hat{x}_i - x_i)^2}{\sigma_{\text{Tech}}^2 + \sigma_{\text{Bio}}^2}, \quad (1)$$

where  $x_i$  is the  $i$ th measurement,  $\hat{x}_i$  is the  $i$ th prediction,  $\sigma_{\text{Tech}}^2$  is the technical variance, and  $\sigma_{\text{Bio}}^2$  is the biological variance. The variances were parametrized as follows:  $\sigma_{\text{Tech}} = 300$  (minimum sensitivity of the detector for antibody-based detection assays) and  $\sigma_{\text{Bio}} = 0.8 \times x_i$  (product of the coefficient of variation and the measurement). Note that the squared prediction error is normalized by an estimate of the measurement variance, a sum of the biological variance and the technical variance. A probability distribution for this metric was estimated by simulation of a null model.

The null model was based on a naive approach to solving the challenge. Essentially, participants were provided a spreadsheet of measurements with some entries

missing. The columns of the spreadsheet corresponded to the phosphoproteins or cytokines (depending on the subchallenge); rows corresponded to various perturbations of stimuli and inhibitors. We randomly “filled-in” the spreadsheet by choosing values for the missing entries, with replacement, from the corresponding column. Since each protein or cytokine had a characteristic dynamic range, this procedure ensured that the random predictions were drawn from the appropriate order of magnitude of fluorescence. This procedure was performed 100,000 times. Parametric curves were fit to the histograms (Figure 2B) to extrapolate the probability density beyond the range of the histogram (i.e., to compute  $p$ -values for teams that did far better or worse than this null model). The procedure for curve-fitting was described previously [10]. Briefly, an approximation of the empirical probability density was given by stretched

exponentials with different parameters to the right and left of the mode of the distribution, with functional form

$$\text{pdf}(x) = \begin{cases} h_{\max} \exp[-b_{>}(x - x_{\max})^{c_{>}}] & \text{for } x \geq x_{\max} \\ h_{\max} \exp[-b_{<}(x - x_{\max})^{c_{<}}] & \text{for } x < x_{\max} \end{cases} \quad (2)$$

where  $h_{\max}$  is the maximum height of the histogram,  $x_{\max}$  is the position of  $h_{\max}$ , and  $b_{>}$ ,  $b_{<}$ ,  $c_{>}$ , and  $c_{<}$  are fitted parameters.

### Gene Expression Prediction

The gene expression prediction challenge explored the extent to which time-dependent gene expression measurements can be predicted in a mutant strain of *S. cerevisiae* (budding yeast) given complete expression data for the wild type strain and two related mutant strains. Experimental perturbations involving histidine biosynthesis provided a context for the challenge. Neil Clarke of the Genome Institute of Singapore generously donated unpublished gene expression measurements for this challenge.

The yeast transcription factors GAT1, GCN4, and LEU3 regulate genes involved in nitrogen and/or amino acid metabolism. They were disrupted in three mutant strains denoted *gat1Δ*, *gcn4Δ*, and *leu3Δ*. These genes are considered nonessential since the deletion strains are viable. Expression levels were assayed separately in the three mutant strains and in the wild type strain at times 0, 10, 20, 30, 45, 60, 90 and 120 minutes following the addition of 3-aminotriazole (3AT) as described [17]. 3AT inhibits an enzyme in the histidine biosynthesis pathway and, in the appropriate media (used in these experiments), has the effect of starving the cells for this essential amino acid.

Expression measurements were obtained using DNA microarrays (Affymetrix YGS98 GeneChip). Two biological replicates (i.e., independent cultures) and an additional technical replicate (i.e., independent labeling and hybridization of the same culture) were performed. Measurements were normalized using the RMA algorithm [18] within the commercial software package, GeneSpring. Values were median normalized within arrays prior to the calculation of fold-change. The mean hybridization value for each probe set was obtained from the three replicates and was normalized to the mean value for the probe set in the wild-type samples at time zero. Values were provided as the log (base two) of the ratio of the indicated experimental condition (i.e., strain and time point) relative to the wild type strain at time zero.



The fifty genes composing the test set were selected by subjective criteria, but with an eye towards enriching for genes that are significantly regulated in at least one strain, or are bound by one or more of the transcription factors according to ChIP-chip data, or are relatively strongly predicted to be bound based on a PWM-based promoter occupancy calculation [19]. Thus, the expression profiles for these genes tend to be somewhat more explicable than would be the case for randomly selected genes. Nevertheless, it was trivial to find genes for which an explanation for the expression profiles was not obvious, and there are many such genes among the fifty prediction targets.

A quantitative prediction of gene expression changes is far beyond state of the art at this time, so participants were asked to predict the relative expression levels for 50 genes at the eight time points in the *gat1Δ* strain. Participants were provided complete expression data for the other strains, as well measurements for the genes that were not part of the set of 50 challenge genes in *gat1Δ*. Predictions for each time point were submitted as a ranked list with values from 1 to 50 sorted from most induced to most repressed compared to the wild type expression at time zero.

**Basis of assessment.** Participants submitted a spreadsheet of 50 rows (genes) by eight columns (time points). Submissions were scored using Spearman's rank correlation coefficient between the predicted and measured gene expression at each of the eight time points. The same statistic was also computed with respect to each gene across all time points. Thus, we evaluated predictions using two different tests of similarity to the gold standard which we call the time-profiles and gene-profiles, respectively.

For each column of the predicted matrix of relative expression, we obtained a correlation coefficient and its corresponding  $p$ -value under the null hypothesis that the ranks are randomly distributed. From the column  $p$ -values we arrived at a single summary  $p$ -value for all eight time points using the geometric mean of individual  $p$ -values (i.e.,  $(p_1 p_2 p_3 p_4 p_5 p_6 p_7 p_8)^{1/8}$ ). The same procedure was performed on the row  $p$ -values to arrive at a summary  $p$ -value for the 50 genes. Finally, the score used to assess best-performers was computed from the two summary  $p$ -values

$$\text{score} = -\frac{1}{2} \log_{10}(p_T \times p_G), \quad (3)$$

where  $p_T$  is the overall  $p$ -value for the time-

profiles (columns) and  $p_G$  is the overall  $p$ -value for the gene-profiles (rows). The higher the score, the more significant the prediction.

### In Silico Network Inference

The *in silico* network inference challenge explored the extent to which gene networks of various sizes and connection densities can be inferred from simulated data. Daniel Marbach of Ecole Polytechnique Fédérale de Lausanne extracted the challenge networks as subgraphs of the currently accepted *E. coli* and *S. cerevisiae* gene regulation networks [8] and imbued the networks with dynamics using a thermodynamic model of gene expression. The *in silico* "measurements" were generated by continuous differential equations which were deemed reasonable approximations of gene expression regulatory functions. To these values was added a small amount of Gaussian noise to simulate measurement error.

The simulated data was meant to mimic three typical types of experiments: (1) time courses of a wild type strain following an environmental perturbation (i.e., trajectories); (2) knock-down of a gene by deletion of one copy in a diploid organism (i.e., heterozygous mutants); (3) knock-out of a gene by deletion of both copies in a diploid organism (i.e., homozygous null mutants). Technically, a haploid organism such as *E. coli* can not be a heterozygote, but since this data only exists *in silico* we did not see harm in the use of this term. A trajectory of the wild-type response to an environmental perturbation was simulated by a random initialization of the simulation. A heterozygous knock-down mutant was simulated by halving the wild type concentration of the gene. A homozygous knock-out mutant was simulated by flooring the wild type concentration of the gene to zero.

The challenge was organized into three parts: the 10-node subchallenge, the 50-node subchallenge, and the 100-node subchallenge. Within each sub-challenge, participants were required to predict five networks, denoted Ecoli1, Ecoli2, Yeast1, Yeast2, Yeast3. Completion of a subchallenge required submission of predictions for all five of the networks in the subchallenge. Participants were encouraged, but not required, to perform all three subchallenges on networks of various sizes. Some of the gross topological properties of the fifteen gold standard networks are illustrated in Table 2.

Complete steady state expression information was provided for the wild type and mutant strains. In other words, in the 10-

node subchallenge, all ten genes were knocked-down and knocked-out, one at a time, while the remaining nine measurements were provided. Various numbers of trajectories from random initializations were provided depending on the subchallenge. Four, 23, and 46 trajectories were provided for the 10-node, 50-node, and 100-node subchallenges, respectively.

Participants were asked to predict the directed, unsigned networks from *in silico* gene expression data sets. A network prediction was submitted in the form of a ranked list of potential network edges ordered from most reliable to least reliable. In other words, the edges at the top of the list were believed to be present in the network and the edges at the bottom of the list were believed to be absent from the network. This submission format was chosen because it does not require the researcher to impose a particular threshold for calling an edge present or absent. Also, it can be scored without imposition of a specific threshold. An example of the file format of a network prediction is illustrated in Table 3.

**Basis of assessment.** From the ranked edge-list (Table 3), a particular concrete network with  $k$  edges is obtained by designating the first  $k$  edges present and the remaining edges absent. Then,  $k$  is a parameter that controls the number of edges in a predicted network. Various performance metrics were computed as  $k$  was varied from 1 to  $T$ , the total number of possible directed edges, where  $T = N(N-1)$  and  $N$  is the number of nodes in the network.

Two parametric curves, the precision-recall (P-R) curve and the receiver operating characteristic (ROC) curve, were traced by scanning  $k=1, \dots, T$ . *Recall* is a measure of completeness,

$$\text{rec}(k) = \frac{\text{TP}(k)}{P},$$

where  $\text{TP}(k)$  is the number of true positives at threshold  $k$ , and  $P$  is the number of positives (i.e., gold standard edges). *Precision* is a measure of fidelity,

$$\text{prec}(k) = \frac{\text{TP}(k)}{\text{TP}(k) + \text{FP}(k)} = \frac{\text{TP}(k)}{k},$$

where  $\text{FP}(k)$  is the number of false positives at threshold  $k$ . Note that the sum of  $\text{TP}(k)$  and  $\text{FP}(k)$  is  $k$ . The precision-recall curve graphically explores the tradeoff between these complementary metrics as the parameter  $k$  is varied. The area under the precision-recall curve

**Table 2.** Statistical properties of the gold standard networks in the *in silico* network inference challenge.

Sub-challenge	Network	Nodes	Edges	Regulators
<b>In Silico Size 10</b>	Ecoli1	10	11	5
	Ecoli2	10	15	3
	Yeast1	10	10	7
	Yeast2	10	25	8
	Yeast3	10	22	9
<b>In Silico Size 50</b>	Ecoli1	50	62	13
	Ecoli2	50	82	11
	Yeast1	50	77	26
	Yeast2	50	160	37
	Yeast3	50	173	35
<b>In Silico Size 100</b>	Ecoli1	100	125	26
	Ecoli2	100	119	19
	Yeast1	100	166	60
	Yeast2	100	389	71
	Yeast3	100	551	81

In each of the three sub-challenges the number of nodes was held constant but the number of edges and regulator nodes was not. There were five gold standard networks in each of the three sub-challenges (which were treated as three separate contests).

doi:10.1371/journal.pone.0009202.t002

(AUPR) is a single number that summarizes the precision-recall tradeoff. Similarly, the receiver operating characteristic (ROC) curve graphically explores the tradeoff between the *true positive rate* (TPR) and the *false positive rate* (FPR).  $TPR(k)$  is the fraction of positives that are correctly predicted at threshold  $k$ ,

$$TPR(k) = \frac{TP(k)}{P}.$$

(Note that TPR is equivalent to recall.)  $FPR(k)$  is the fraction of negatives that are

incorrectly predicted at threshold  $k$ ,

$$FPR(k) = \frac{FP(k)}{N}.$$

*Negative* denotes the absence of an edge in the gold standard network. The area under the ROC curve (AUROC, also denoted AUC in the literature) is a single number that summarizes the tradeoff between  $TPR(k)$  and  $FPR(k)$  as the parameter  $k$  is varied. Using both the AUPR and the AUROC metrics, we gain a fuller characterization of the prediction

than using either alone. For example, the P-R curve indicates whether the first few edge predictions at the top of the prediction list are correct. The ROC curve does not provide this information.

A technical point is the issue of how to score a truncated prediction list, where fewer than the total number of possible edges are submitted. A methodology is in place from the previous DREAM assessment [10]. If a prediction list does not contain a complete ordering of all possible  $N(N-1)$  edges, we “add” the missing edges in random order at the end of the list. The addition takes place in an analytical way.

A team’s score for a subchallenge depended on quite a few calculations. Each of the five network predictions (Ecoli1, Ecoli2, Yeast1, Yeast2, Yeast3) were evaluated by AUPR and AUROC.  $P$ -values for these assessments were obtained from the empirical distributions described above. The five AUPR  $p$ -values were condensed to an overall AUPR  $p$ -value using the geometric mean of individual  $p$ -values (i.e.,  $(p_1 p_2 p_3 p_4 p_5)^{1/5}$ ). The same procedure was performed on the five AUROC  $p$ -values to arrive at an overall AUROC  $p$ -value. Finally, the score for the team was computed as

$$\text{score} = -\frac{1}{2} \log_{10}(p_{AUPR} \times p_{AUROC}), \quad (4)$$

where  $p_{AUPR}$  and  $p_{AUROC}$  are the overall  $p$ -values for AUPR and AUROC, respectively. The higher the score, the more significant the network prediction.

## Results and Discussion

The DREAM3 challenges were posted on the DREAM website on June 15, 2008. Submissions in response to the challenges were accepted on September 15, 2008. Forty teams submitted 413 predicted networks and test set predictions in the various challenges. The anonymous results were posted on the DREAM website [12] on October 15, 2008.

In this section, we describe our assessment of the predictions supplied by the community. Our dual goals are to identify the best-performers in each challenge and to characterize the efficacy of the community as a whole. We highlight the best-performer strategies and comment on some of the sub-optimal strategies. Where possible, we attempt to leverage the community intelligence by combining the predictions of multiple teams into a consensus prediction.

Best-performers in each challenge were identified by statistical significance with respect to a null model combined with a

**Table 3.** Format of a predicted network in the *in silico* network inference challenge.

Source Node	Target Node	Confidence	Scoring Cutoff (k)
G85	G1	1.00	1
G85	G10	0.99	2
G10	G85	0.73	3
G99	G52	0.44	4
⋮	⋮	⋮	⋮
G10	G3	0.01	N(N-1)

Predicted edges were to be ranked from most confidence to least confidence that the edge is present in the network. A directed edge is denoted by a source and target node and an arbitrary (non-increasing) score between one (most confidence) to zero (least confidence). Thus, edges that are predicted to exist in the network should be at the top of the list and those predicted not to exist in the network should be at the bottom of the list. To evaluate the predicted network, two metrics—area under the ROC curve and area under the precision-recall curve—were computed by scanning all possible decision boundaries (i.e.,  $k = 1$ ,  $k = 2$ , etc.) up to the maximum number of possible directed edges (excluding self-edges).

doi:10.1371/journal.pone.0009202.t003

clear delineation from the rest of the participating teams (e.g., an order of magnitude lower  $p$ -value compared to the next best team). Occasionally, this criterion identified multiple best-performers in a challenge.

### Signaling Cascade Identification

Seven teams submitted predictions for the signaling cascade identification challenge as described in the Introduction. Submissions were scored based on the probability that a random solution to the challenge would achieve at least as many correct protein identifications as the submitted solution.

Five of seven teams identified two of the four proteins correctly (though not the same pair) (Table 4). One team identified only one protein correctly and one team did not identify any correctly. The  $p$ -value for a team identifying two or more proteins correctly is 0.11, as described in the Introduction. On the basis of this  $p$ -value, this challenge did not have a best-performer. However, in the days following the conference, follow-up questions from some of the participants to the data provider revealed a misrepresentation in how the challenge was posed, which probably negatively impacted the teams' performances. The source of the confusion is describe below.

Despite that no individual team gained much traction in solving this challenge, the community as a whole seemed to possess intelligence. For example, five of seven teams correctly identified two proteins (though not the same pair). While such a performance is not significant on an individual basis, the event of five teams correctly identifying two proteins is unlikely to occur by chance. Under the

binomial distribution, assuming independent teams, the probability of five or more teams correctly identify two or more proteins is  $2.6 \times 10^{-4}$ .

Summing over the predictions of all the teams we obtain Figure 3. For example, five of seven teams correctly identified x1 as the kinase. The probability that five or more teams would pick the same table entry is  $9.7 \times 10^{-4}$ . Similarly, the probability of three or more teams identifying the same pair of proteins (e.g., kinase, phosphoprotein) is  $4.4 \times 10^{-4}$ .

The assumption of independence is implicit in the null hypothesis underlying these  $p$ -values. Rejection of the null hypothesis on the basis of a small  $p$ -value indicates that there is a correlation between the teams. This correlation can be interpreted as a shared success within the community. In other words, the community exhibits some intelligence not evidenced in the predictions of the individual teams. Based on this assessment of the community as a whole, we conclude that some structural features of the signaling cascade were indeed identified from flow cytometry data.

The community assessment suggests that a mixture of methods may be an advantageous strategy for identifying signaling proteins from flow-cytometry data. A simple strategy for generating a consensus prediction is illustrated by Figure 3 in which the total number of predictions made by the community for each possible assignment are indicated along with the corresponding  $p$ -values indicating the probability of such a concentration of predictions in a single table entry. The kinase and phosphorylated protein are the only identifications (individually) significant at  $p < 0.05$ . This analysis also reveals clustering of incorrect predictions—the phosphatase was most often confused with the activated phosphatase, and the phosphorylated protein was most often confused with the phosphorylated ligand-receptor complex—but these misidentifications were not significant.

#### Mea culpa: a poorly posed challenge.

There are three conjugate pairs of species in the signaling pathway: complex/phospho-complex, protein/phospho-protein, and phosphatase/activated phosphatase. The challenge description led participants to believe that each measured species (x1,..., x4) may match one of the six individual species. In fact, measurement x3 corresponded to total protein (inactive and active forms). Likewise, measurement x2 corresponded to total phosphatase (inactive and active forms). It would be highly unusual for an antibody to target one

epitope of a protein to the exclusion of a phosphorylated epitope. That is, it would be difficult but not impossible to raise an antibody that reacted with only the unphosphorylated version of a protein. This serious flaw in the design of the challenge did not come to light until after the scoring was complete.

The simultaneous identification of the upstream kinase and the downstream phosphorylated protein (Figure 3) can be explained in light of the confusion surrounding precisely what the measurements entailed. The measurements corresponding to the kinase and phosphoprotein were accurately portrayed in the challenge description whereas the total protein and total phosphatase were not.

### Signaling Response Prediction

Four teams participated in the signaling response prediction challenge. The phosphoprotein subchallenge received three submissions, as did the cytokine subchallenge. As described in the Introduction, the task was to predict measurements of proteins and/or cytokines, in normal and cancerous cells, for combinatoric perturbations of stimuli and inhibitors of a signaling pathway. Submissions were scored by a metric based on the sum of the squared prediction errors (Figure 2B).

In the phosphoprotein subchallenge two teams achieved a  $p$ -value orders of magnitude lower than the remaining other submission (Table 5). In the cytokine subchallenge one team had a substantially smaller total prediction error than the next best team. On this basis, the best-performers were:

- Genome Singapore (phosphoprotein and cytokine subchallenges): Guillaume Bourque and Neil Clarke of the Genome Institute of Singapore, Singapore
- Vital SIB (phosphoprotein subchallenge): Nicolas Guex, Eugenia Miglavacca, and Ioannis Xenarios of the Swiss Institute of Bioinformatics, Switzerland

There are two main types of strategies that could have been employed in this challenge: to explicitly model the underlying signaling network, or to model the data statistically. Both of the best-performers took a statistical approach. Vital SIB approached it as a missing data problem and used multiple imputation to predict the missing data. This involved learning model parameters by cross-validation, followed by prediction of the missing data [20]. Genome Singapore identified the

**Table 4.** Results of the signaling cascade identification challenge.

Team	No. Correct	
	Identifications	$p$ -value
Team 315	2	0.11
Team 283	2	0.11
Team 106	2	0.11
Team 281	2	0.11
Team 181	2	0.11
Team 110	1	0.45
Team 286	0	1.00

Two correct identifications is not a significant performance so no team was named the best-performer.

doi:10.1371/journal.pone.0009202.t004



Measurement Label	Kinase	Protein	Phospho-protein	Phosphatase	Activated Phosphatase	Complex	Phospho-complex
x1	5 (.001)					2 (.3)	
x2				2 (.3)	3 (.07)		2 (.3)
x3		0 (1)	2 (.3)		2 (.3)	1 (.7)	2 (.3)
x4			4 (.01)				3 (.07)

**Figure 3. Overlay of the assignment tables from the seven teams in the signaling cascade identification challenge.** The number of teams making each assignment and the  $p$ -value is indicated. The  $p$ -value expresses the probability of a such a concentration of random guesses in the same table entry. Highlighted entries are correct. Five teams correctly identified species x1 as the kinase, a significant event for the community despite that no team had a significant individual performance.

doi:10.1371/journal.pone.0009202.g003

nearest-neighbors of missing measurements based on similarity of the measurement profiles [21]. To predict the measurements for an unobserved stimulus or inhibitor, they took into consideration the values observed for the nearest neighbor. Neither team utilized external data sources, nor did they evoke the concept of a biological signaling network.

Surprisingly, one team in the cytokine subchallenge had a significantly larger total error than random. We investigated this strange outcome further. This team systematically under-predicted the medium and large intensity measurements (data not shown). This kind of systematic error was heavily penalized by the scoring metric. Nevertheless, the best-performer would have remained the same had linear correlation been used as the metric. Due to the low participation level from the community, we did not perform a community-wide analysis.

### Gene Expression Prediction

Nine teams participated in the gene expression prediction challenge as described in the Introduction. The task was to predict the expression of 50 genes in the *gat1Δ* strain of *S. cerevisiae* at eight time points. Participants submitted a spreadsheet of 50 rows (genes) by eight columns (time points). At each time point, the participant ranked the genes from most induced to most repressed compared to the wild type values at time zero. Predictions

were assessed by Spearman's correlation coefficient and its corresponding  $p$ -value under the null hypothesis that the ranks are uniformly distributed.

The  $p$ -values (based on Spearman correlation coefficient) computed over the set of 50 test genes at each of the eight time-points are reported in Table 6. Some trends are readily identifiable. Across the community, the least significant predictions were those at time zero. Relatively more significant predictions were made at 10, 20, 45, and 60 minutes, and comparatively less significant predictions were made at 30 and 90 minutes. This analysis identified the teams that predicted well (over the 50 test genes) at each time point. We computed a summary statistic for each team using the geometric mean of the eight  $p$ -values for the individual time points.

In the above analysis, each of the eight time points was analyzed as a 50-dimensional vector. An alternative viewpoint is to consider each of the 50 genes as an eight-dimensional vector. We also performed this analysis using Spearman's correlation coefficient computed for each gene. We computed a summary statistic for each team using the geometric mean of the 50  $p$ -values for the individual genes (not shown). Correlation coefficients and  $p$ -values for the gene-profiles are published on the DREAM website [12].

Summary statistics from the time-profile analysis and the gene-profile analysis are

reported in Table 7. Weaker significance of gene-profile  $p$ -values compared to time-profile  $p$ -values may be due to the fact that the former are eight-dimensional vectors while the latter are 50-dimensional vectors. Best-performers were identified by an overall score based on the time-profile and gene-profile summary  $p$ -values. A difference of one in the overall score corresponds to an order of magnitude difference in the  $p$ -value. Two teams performed more than an order of magnitude better than the nearest competitor at  $p < 0.05$ .

- Gustafsson-Hornquist : Mika Gustafsson and Michael Hornquist of Linköping University, Sweden
- Dream Team 2008 : Jianhua Ruan of the University of Texas at San Antonio, USA

We used hierarchically clustered heat maps to visualize the teams' predictions (gene ranks from 1 to 50) relative to the gold standard (Figure 4A). The two best-performers were more similar to each other than either was to the gold standard. The Spearman correlation coefficient between Gustafsson-Hornquist and Dream Team 2008 is 0.96, while the correlation between either team and the Gold Standard is 0.67. One could reasonably presume that substantially similar methods were employed by both teams. That turns out not to be the case.

Team Gustafsson-Hornquist used a weighted least squares approach in which the prediction for each gene was a weighted sum of the values of the other genes [22]. The particular linear model they employed is called an elastic net, which is a hybrid of the lasso and ridge regression [23]. They incorporated additional data into their model, taking advantage of public yeast expression profiles and ChIP-chip data. The additional expression profiles provided more training examples from which to estimate pairwise correlations between genes. The physical binding data (ChIP-chip) was integrated into the linear model by weighting each gene's contribution to a

**Table 5. Results of the signaling response prediction challenge.**

Subchallenge	Team	Norm. Sq. Error	$p$ -value
Phosphoprotein	Vital SIB	3102	$2 \cdot 10^{-22}$
	GenomeSingapore	3310	$4 \cdot 10^{-22}$
	Team 302	11329	$7 \cdot 10^{-14}$
Cytokine	GenomeSingapore	4462	$8 \cdot 10^{-36}$
	Team 302	13995	$4 \cdot 10^{-9}$
	Team 126	29795	1

doi:10.1371/journal.pone.0009202.t005

**Table 6.** Time-profile  $p$ -values of the gene expression prediction challenge.

	Minutes								Summary
	0	10	20	30	45	60	90	120	Time-profile
Team	$p$ -value	$p$ -value	$p$ -value	$p$ -value	$p$ -value	$p$ -value	$p$ -value	$p$ -value	$p$ -value
Gustafsson-Hornquist	$2 \cdot 10^{-2}$	$1 \cdot 10^{-7}$	$3 \cdot 10^{-7}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-7}$	$1 \cdot 10^{-7}$	$7 \cdot 10^{-4}$	$3 \cdot 10^{-6}$	$6.5 \cdot 10^{-6}$
Dream Team 2008	$3 \cdot 10^{-3}$	$4 \cdot 10^{-7}$	$3 \cdot 10^{-6}$	$5 \cdot 10^{-4}$	$1 \cdot 10^{-6}$	$1 \cdot 10^{-7}$	$2 \cdot 10^{-4}$	$6 \cdot 10^{-6}$	$1.1 \cdot 10^{-5}$
Team 263	$7 \cdot 10^{-2}$	$1 \cdot 10^{-4}$	$7 \cdot 10^{-6}$	$4 \cdot 10^{-3}$	$2 \cdot 10^{-3}$	$2 \cdot 10^{-2}$	$8 \cdot 10^{-4}$	$2 \cdot 10^{-5}$	$7.5 \cdot 10^{-4}$
Team 297	$8 \cdot 10^{-2}$	$1 \cdot 10^{-2}$	$7 \cdot 10^{-5}$	$4 \cdot 10^{-4}$	$1 \cdot 10^{-3}$	$4 \cdot 10^{-3}$	$4 \cdot 10^{-2}$	$3 \cdot 10^{-1}$	$5.6 \cdot 10^{-3}$
Team 126	$1 \cdot 10^{-1}$	$5 \cdot 10^{-3}$	$8 \cdot 10^{-3}$	$2 \cdot 10^{-2}$	$8 \cdot 10^{-4}$	$1 \cdot 10^{-3}$	$9 \cdot 10^{-2}$	$8 \cdot 10^{-3}$	$9.0 \cdot 10^{-3}$
Team 273	$3 \cdot 10^{-1}$	$1 \cdot 10^{-2}$	$4 \cdot 10^{-3}$	1	$1 \cdot 10^{-5}$	$6 \cdot 10^{-3}$	$9 \cdot 10^{-2}$	$2 \cdot 10^{-5}$	$6.1 \cdot 10^{-3}$
Team 186	$1 \cdot 10^{-1}$	$2 \cdot 10^{-2}$	$7 \cdot 10^{-1}$	$7 \cdot 10^{-1}$	$2 \cdot 10^{-1}$	$9 \cdot 10^{-2}$	$2 \cdot 10^{-3}$	$1 \cdot 10^{-4}$	$3.9 \cdot 10^{-2}$
Team 190	$3 \cdot 10^{-2}$	$1 \cdot 10^{-1}$	$6 \cdot 10^{-3}$	$2 \cdot 10^{-1}$	$6 \cdot 10^{-2}$	$3 \cdot 10^{-2}$	$2 \cdot 10^{-2}$	$7 \cdot 10^{-2}$	$4.2 \cdot 10^{-2}$
Team 193	$1 \cdot 10^{-1}$	1	1	$8 \cdot 10^{-1}$	1	1	1	1	$7.4 \cdot 10^{-1}$

$P$ -values at each time-point and a summary  $p$ -value (geometric mean) are indicated.  
doi:10.1371/journal.pone.0009202.t006

prediction based on the number of common transcription factors the pair of genes shared.

Dream Team 2008 did not use any additional data beyond what was provided in the challenge. Rather, they employed a  $k$ -nearest neighbor (KNN) approach to predict the expression of a gene based on the expression of other genes in the same strain at the same time point [24]. The Euclidean distance between all pairs of genes was determined from the strains for which complete expression profiles were provided. The predicted value of a gene was the mean expression of the  $k$ -nearest-neighbors. The parameter  $k$  was chosen by cross-validation;  $k=10$  was used for prediction.

Does the community possess an intelligence that trumps the efforts of any single

team? To answer this question we created a consensus prediction by summing the predictions of multiple teams, then re-ranking. The results of this analysis are shown in Figure 4B which traces the overall score of the consensus prediction as lower-significance teams are included. The first consensus prediction includes the best and second-best teams. The next consensus prediction includes the top three teams, and so on.

The consensus prediction of the top four teams had a higher score than the best-performer, which is counter-intuitive since the third and fourth place teams individually scored much lower than the best-performer (Figure 4B). Furthermore, the inclusion of all teams in the consensus prediction scored about the same as the best-performer. This result suggests that,

given the output of a collection of algorithms, combining multiple result sets into a consensus prediction is an effective strategy for improving the results.

We assigned a difficulty level to each gene based on the accuracy of the community. For each gene, we computed the geometric mean of the gene-profile  $p$ -values over the nine teams, which we interpreted as the difficulty level of each gene. The five best-predicted genes were: *arg4*, *ggc1*, *tmt1*, *arg1*, and *arg3*. The five worst-predicted genes were: *srx1*, *lee1*, *sol4*, *glo4*, and *bap2*. The relative difficulty of prediction of a gene was weakly correlated with the absolute expression level of that gene at  $t=0$ , but many of the 50 genes defied a clear trend. The five best-predicted genes had an average expression of 42.7 (arbitrary units, log scale) at  $t=0$ , whereas the five worst-predicted genes had an average expression of 3.7. It is known that low intensity signals are more difficult to characterize with respect to the noise. It is likely that the absolute intensity of the genes played a role in the relative difficulty of predicting their expression values.

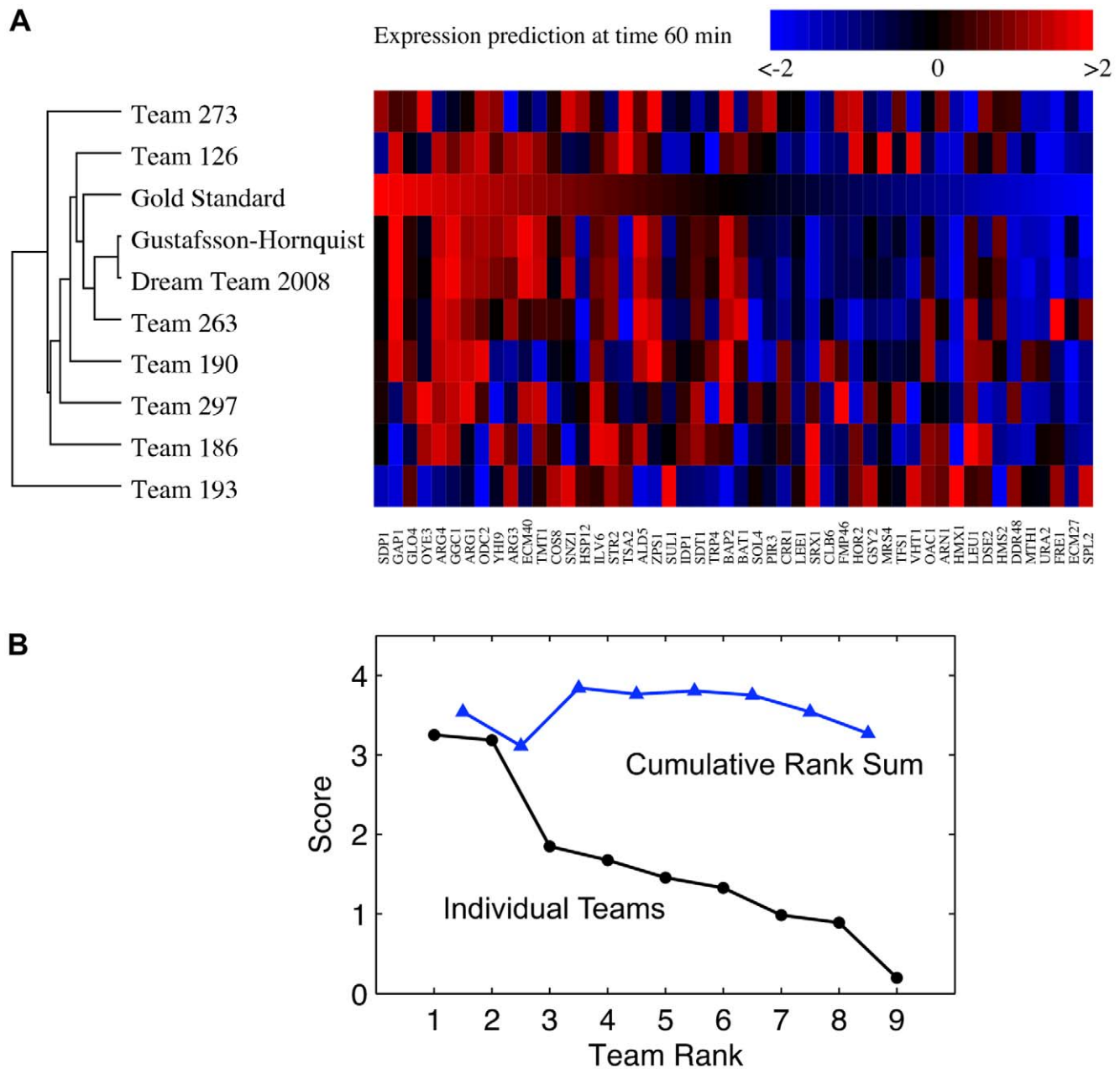
### In Silico Network Inference

Twenty-nine teams participated in the *in silico* network inference challenge as described in the Introduction, the greatest level of participation by far of the four DREAM3 challenges. The task was to infer the underlying gene regulation networks from *in silico* measurements of environmental perturbations (dynamic trajectories), gene knock-downs (heterozygous mutants), and gene knock-outs (homozygous null-mutants). Participants

**Table 7.** Results of the gene expression prediction challenge.

Team	Time-profile	Gene-profile	Overall
	$p$ -value	$p$ -value	Score
Gustafsson-Hornquist	$7 \cdot 10^{-6}$	$5 \cdot 10^{-2}$	3.3
Dream Team 2008	$1 \cdot 10^{-5}$	$4 \cdot 10^{-2}$	3.2
Team 263	$8 \cdot 10^{-4}$	$3 \cdot 10^{-1}$	1.8
Team 297	$6 \cdot 10^{-3}$	$8 \cdot 10^{-2}$	1.7
Team 126	$9 \cdot 10^{-3}$	$1 \cdot 10^{-1}$	1.5
Team 273	$6 \cdot 10^{-3}$	$4 \cdot 10^{-1}$	1.3
Team 186	$4 \cdot 10^{-2}$	$3 \cdot 10^{-1}$	1.0
Team 190	$4 \cdot 10^{-2}$	$4 \cdot 10^{-1}$	0.9
Team 193	$7 \cdot 10^{-1}$	$5 \cdot 10^{-1}$	0.2

doi:10.1371/journal.pone.0009202.t007



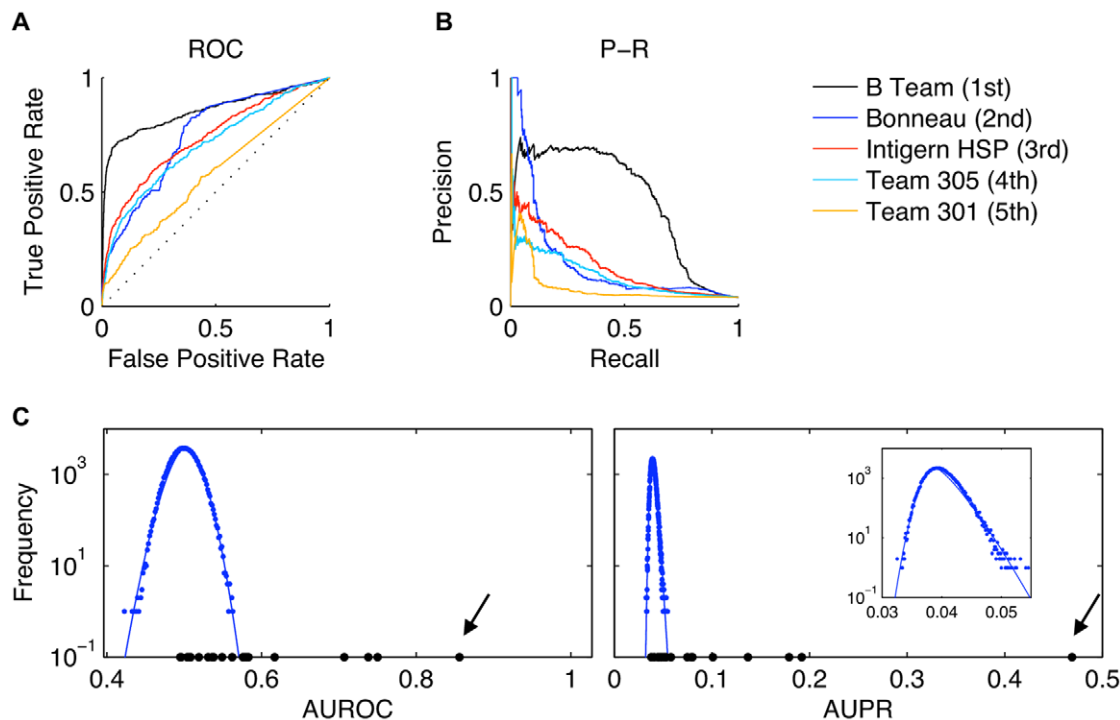
**Figure 4. The objective of the gene expression prediction challenge was to predict temporal expression of 50 genes that were withheld from a training set consisting of 9285 genes. (a)** Clustered heatmaps of the predicted genes (columns) reveal that two best-performer teams predicted substantially similar gene expression values, though different methods were employed. Results for the 60 minute time-point are shown. **(b)** The benefits of combining the predictions of multiple teams into a consensus prediction are illustrated by the rank sum prediction (triangles). Some rank sum predictions score higher than the best-performer, depending on the teams that are included. The highest score is achieved by a combination of the predictions of the best four teams.

doi:10.1371/journal.pone.0009202.g004

predicted directed, unsigned networks as a ranked list of potential edges in order of the confidence that the edge is present in the gold standard network. Predictions for 15 different networks of various “real-world” inspired topologies were solicited, grouped into three separate subchallenges: the 10-node, 50-node, and 100-node subchallenges. The three subchallenges were evaluated separately.

Each predicted network was evaluated using two metrics, the area under the ROC curve (AUROC) and the area under the precision-recall curve (AUPR). To provide some context for these metrics we demonstrate the ROC and P-R curves for the five best teams in the 100-node subchallenge (Figure 5A, 5B). These complementary assessments enable valuable insights about the performance of the various teams.

Based on the P-R curve, we observe that the best-performer in this subchallenge actually had low precision at the top of the prediction list (i.e., the first few edge predictions were false positives), but subsequently maintained a high precision (approximately 0.7) to considerable depth in the prediction list. By contrast, the second-place team had perfect precision for the first few predictions, but precision



**Figure 5. The objective of the *in silico* network inference challenge was to infer networks of various sizes (10, 50, and 100 nodes) from steady-state and time-series “measurements” of simulated gene regulation networks.** Predicted networks were evaluated on the basis of two scoring metrics, (a) area under the ROC curve and (b) area under the precision-recall curve. ROC and precision-recall curves of the five best teams in the 100-node sub-challenge. (a) Dotted diagonal line is the expected value of a random prediction. (b) Note that the best and second-best performers have different precision-recall characteristics. (c) Histograms (log scale) of the AUROC scoring metric for 100,000 random predictions was approximately Gaussian (fitted blue points) whereas the histogram of the AUPR metric was not (inset). Significance of the predictions of the teams (black points) was assessed with respect to the empirical probability densities embodied by these histograms. Scores of the best-performer team are denoted with arrows. All plots are analyses of the gold standard network called *InSilico\_Size100\_Yeast2*. doi:10.1371/journal.pone.0009202.g005

then plummeted. In another example of the complementary nature of the two assessments, consider the fifth-place team. On the basis of the ROC, the fifth place team is scarcely better than random (diagonal dotted line) however, on the basis of the P-R curve, it is clear that the fifth place team achieved better precision than random at the top of edge list. The two types of curves are non-redundant and enable a fuller characterization of prediction performance than either alone.

ROC and P-R curves like those shown in Figure 5 were summarized using the area under the curve. The details of the calculation of the area under the ROC curve and the area under the P-R curve are described at length in [10]. Probability densities for AUPR and AUROC were estimated by simulation of 100,000 random prediction lists. Curves were fit to the histograms using Equation 2 so that the probability densities could be extrapolated beyond the ranges of the histograms in order to compute *p*-values for teams that predicted much better or worse than the null model. Figure 5C demonstrates the teams' scores in the reconstruction of the

gold standard network called *InSilico\_Size100\_Yeast2*. The best-performer made an exceedingly significant network prediction (identified by an arrow) whereas many of the teams predicted equivalently to random.

Best-performers in each subchallenge were identified by an overall score that summarized the statistical significance of the five network reconstructions composing the subchallenge (*Ecoli1*, *Ecoli2*, *Yeast1*, *Yeast2*, *Yeast3*). The AUROC *p*-values for the 100-node subchallenge are indicated in Table 8. The complete set of tables for the other subchallenges are available on the DREAM website [12]. A summary *p*-value for AUROC was computed as the geometric mean of the five *p*-values. Likewise, a summary *p*-value for AUPR was computed (not shown). Finally, the overall score for a team was computed from the two summary *p*-values according to Equation 4 (Table 9). A difference of one in the score corresponds to an order of magnitude difference in *p*-value—the higher the score, the more significant the prediction. On the basis of the overall score, the same team was the

best-performer in the 10-node, 50-node, and 100-node subchallenges:

- B Team : Kevin Y. Yip, Roger P. Alexander, Koon-Kiu Yan, and Mark Gerstein of Yale University, USA

Runners-up were identified by scores that were orders of magnitude more significant than the community at large, but not as significant as the best-performer:

- USMtec347 (10-node, 50-node): Peng Li and Chaoyang Zhang of the University of Southern Mississippi, USA
- Bonneau (100-node): Aviv Madar, Alex Greenfield, Eric Vanden-Eijnden, and Richard Bonneau of New York University, USA
- Intigern HSP (100-node): Xuebing Wu, Feng Zeng, and Rui Jiang of Tsinghua University, China

The overall *p*-values for the 100-node subchallenge (Table 9) demonstrates that the best teams predicted significantly better than the null model—a randomly sorted prediction list. However, the ma-

**Table 8.** *P*-values for the area under the ROC in the *in silico* size 100 network inference challenge.

	Ecoli1	Ecoli2	Yeast1	Yeast2	Yeast3	Summary
Team	<i>p</i> -value	<i>p</i> -value	<i>p</i> -value	<i>p</i> -value	<i>p</i> -value	<i>p</i> -value
B Team	$1 \cdot 10^{-52}$	$6 \cdot 10^{-42}$	$4 \cdot 10^{-70}$	$6 \cdot 10^{-99}$	$2 \cdot 10^{-92}$	$3 \cdot 10^{-71}$
Bonneau	$1 \cdot 10^{-34}$	$5 \cdot 10^{-26}$	$1 \cdot 10^{-33}$	$6 \cdot 10^{-52}$	$9 \cdot 10^{-35}$	$3 \cdot 10^{-36}$
IntigernHSP	$3 \cdot 10^{-31}$	$2 \cdot 10^{-30}$	$3 \cdot 10^{-47}$	$3 \cdot 10^{-47}$	$6 \cdot 10^{-35}$	$3 \cdot 10^{-38}$
Team 305	$1 \cdot 10^{-29}$	$4 \cdot 10^{-27}$	$8 \cdot 10^{-43}$	$7 \cdot 10^{-37}$	$2 \cdot 10^{-30}$	$2 \cdot 10^{-33}$
Team 301	$1 \cdot 10^{-6}$	$3 \cdot 10^{-8}$	$4 \cdot 10^{-13}$	$4 \cdot 10^{-7}$	$9 \cdot 10^{-8}$	$1 \cdot 10^{-8}$
Team 310	$8 \cdot 10^{-5}$	$1 \cdot 10^{-3}$	$5 \cdot 10^{-7}$	$3 \cdot 10^{-5}$	$1 \cdot 10^{-2}$	$1 \cdot 10^{-4}$
Team 314	$4 \cdot 10^{-7}$	$2 \cdot 10^{-1}$	$5 \cdot 10^{-6}$	$7 \cdot 10^{-4}$	$1 \cdot 10^{-5}$	$8 \cdot 10^{-5}$
Team 183	$4 \cdot 10^{-12}$	$3 \cdot 10^{-8}$	$3 \cdot 10^{-14}$	$1 \cdot 10^{-13}$	$9 \cdot 10^{-16}$	$8 \cdot 10^{-13}$
Team 254	$4 \cdot 10^{-6}$	$3 \cdot 10^{-5}$	$5 \cdot 10^{-14}$	$5 \cdot 10^{-8}$	$3 \cdot 10^{-10}$	$1 \cdot 10^{-8}$
Team 192	$3 \cdot 10^{-1}$	$2 \cdot 10^{-2}$	$1 \cdot 10^{-1}$	$1 \cdot 10^{-1}$	$2 \cdot 10^{-1}$	$1 \cdot 10^{-1}$
Team 110	$8 \cdot 10^{-3}$	$3 \cdot 10^{-2}$	$2 \cdot 10^{-7}$	$2 \cdot 10^{-7}$	$3 \cdot 10^{-1}$	$3 \cdot 10^{-4}$
Team 303	$5 \cdot 10^{-3}$	$1 \cdot 10^{-5}$	$5 \cdot 10^{-1}$	$1 \cdot 10^{-7}$	1	$1 \cdot 10^{-3}$
Team 236	$1 \cdot 10^{-1}$	$4 \cdot 10^{-2}$	$8 \cdot 10^{-2}$	$3 \cdot 10^{-1}$	$3 \cdot 10^{-1}$	$1 \cdot 10^{-1}$
Team 283	$6 \cdot 10^{-6}$	$2 \cdot 10^{-4}$	$3 \cdot 10^{-1}$	$4 \cdot 10^{-1}$	$4 \cdot 10^{-1}$	$9 \cdot 10^{-3}$
Team 291	$2 \cdot 10^{-1}$	$5 \cdot 10^{-2}$	1	$5 \cdot 10^{-7}$	1	$2 \cdot 10^{-2}$
Team 271	$3 \cdot 10^{-4}$	$4 \cdot 10^{-3}$	$3 \cdot 10^{-2}$	$1 \cdot 10^{-2}$	$2 \cdot 10^{-1}$	$1 \cdot 10^{-2}$
Team 273	$2 \cdot 10^{-1}$	$1 \cdot 10^{-7}$	$8 \cdot 10^{-1}$	$2 \cdot 10^{-2}$	$1 \cdot 10^{-1}$	$8 \cdot 10^{-3}$
Team 70	$4 \cdot 10^{-1}$	$8 \cdot 10^{-4}$	$4 \cdot 10^{-1}$	$3 \cdot 10^{-1}$	$1 \cdot 10^{-2}$	$6 \cdot 10^{-2}$
Team 302	$2 \cdot 10^{-1}$	$5 \cdot 10^{-2}$	$8 \cdot 10^{-1}$	$7 \cdot 10^{-3}$	$5 \cdot 10^{-3}$	$5 \cdot 10^{-2}$
Team 269	$2 \cdot 10^{-1}$	$6 \cdot 10^{-1}$	$2 \cdot 10^{-1}$	$4 \cdot 10^{-1}$	$5 \cdot 10^{-1}$	$3 \cdot 10^{-1}$
Team 282	$6 \cdot 10^{-1}$	$4 \cdot 10^{-1}$	$5 \cdot 10^{-1}$	$6 \cdot 10^{-1}$	$6 \cdot 10^{-1}$	$5 \cdot 10^{-1}$
Team 280	$6 \cdot 10^{-1}$	$6 \cdot 10^{-1}$	$6 \cdot 10^{-1}$	$7 \cdot 10^{-1}$	$7 \cdot 10^{-1}$	$6 \cdot 10^{-1}$

*P*-values for the area under the ROC curve for each of the five networks in the size-100 sub-challenge and a summary *p*-value (geometric mean) are indicated. The table is sorted in the same order as Table 9.

doi:10.1371/journal.pone.0009202.t008

jority of teams did not predict much better than the null model. In the 10-node subchallenge, twenty-six of twenty-nine teams did not make statistically significant predictions on the basis of the AUROC ( $p < 0.01$ ). Fourteen of 27 teams in the 50-node subchallenge did not make significant predictions (AUROC  $p < 0.01$ ). Eight of 22 teams in the 100-node subchallenge did not make significant predictions (AUROC  $p < 0.01$ ). This is a sobering result for the efficacy of the network inference community. In Conclusions we discuss some reasons for this seemingly distressing result.

Some teams' methods were well-suited to smaller networks, others to larger networks (Table 10). This may have less to do with the number of nodes and more to do with the relative sparsity of the larger networks since the number of potential edges grows geometrically with the number of nodes (i.e.,  $N(N-1)$ ).

B Team used a collection of unsupervised methods to model both the genetic perturbation data (steady-state) and the

dynamic trajectories [25]. Most notably, they correctly assumed an appropriate noise model (additive noise), and characterized changes in gene expression relative to the typical variance observed for each gene. It turned-out that this simple treatment of measurement noise was credited with their overall exemplary performance. This conclusion is based on our own ability to recapitulate their performance using a simple method that also uses a noise model to infer connections (see analysis of null-mutant  $\mathcal{Z}$ -scores below). Additionally, B Team employed a few formulations of ODEs (linear functions, sigmoidal functions, etc.) to model the dynamic trajectories. In retrospect, their efforts to model the dynamic trajectories probably had a minor effect on their overall performance. Team Bonneau applied and extended a previously described algorithm, the Inferelator [26], which uses regression and variable selection to identify transcriptional influences on genes [27]. The methodologies of B Team and the other best-performers are described in

separate publications in the PLoS ONE Collection.

**A simple method: null-mutant  $z$ -score.** We investigated the utility of a very simple network inference strategy which we call the null-mutant  $z$ -score. This strategy is a simplification of conditional correlation analysis [28]. Suppose there is a regulatory interaction which we denote  $A \rightarrow B$ . We assume that a large expression change in B occurs when A is deleted (compared to the wild-type expression). We compute the  $z$ -score for the regulatory interaction  $A \rightarrow B$

$$z_{A \rightarrow B} = \frac{x_{B, \Delta A} - \mu_B}{\sigma_B},$$

where  $x_{B, \Delta A}$  is the value of B in the strain in which A was deleted,  $\mu_B$  is the mean value of B in all strains (WT and mutants), and  $\sigma_B$  is the standard deviation of B in all strains. This calculation is performed for all directed pairs (A, B). We assume that  $\mu_B$  represents baseline expression (i.e., most gene deletions do not affect



**Table 9.** Results of the *in silico* size 100 network inference challenge.

Team	AUROC	AUPR	Overall
	<i>p</i> -value	<i>p</i> -value	Score
B Team	$3 \times 10^{-71}$	0	Inf*
Bonneau	$3 \times 10^{-36}$	$4 \times 10^{-56}$	45.44
IntigernHSP	$3 \times 10^{-38}$	$1 \times 10^{-47}$	42.24
Team 305	$2 \times 10^{-33}$	$8 \times 10^{-32}$	31.88
Team 301	$1 \times 10^{-8}$	$8 \times 10^{-17}$	11.99
Team 310	$1 \times 10^{-4}$	$2 \times 10^{-14}$	8.83
Team 314	$8 \times 10^{-5}$	$3 \times 10^{-14}$	8.81
Team 183	$8 \times 10^{-13}$	$4 \times 10^{-5}$	8.25
Team 254	$1 \times 10^{-8}$	$2 \times 10^{-8}$	7.83
Team 192	$1 \times 10^{-1}$	$8 \times 10^{-8}$	4.05
Team 110	$3 \times 10^{-4}$	$9 \times 10^{-3}$	2.79
Team 303	$1 \times 10^{-3}$	$5 \times 10^{-3}$	2.61
Team 236	$1 \times 10^{-1}$	$3 \times 10^{-4}$	2.21
Team 283	$9 \times 10^{-3}$	$7 \times 10^{-3}$	2.12
Team 291	$2 \times 10^{-2}$	$5 \times 10^{-3}$	1.99
Team 271	$1 \times 10^{-2}$	$3 \times 10^{-2}$	1.75
Team 273	$8 \times 10^{-3}$	$1 \times 10^{-1}$	1.53
Team 70	$6 \times 10^{-2}$	$3 \times 10^{-2}$	1.41
Team 302	$5 \times 10^{-2}$	$1 \times 10^{-1}$	1.17
Team 269	$3 \times 10^{-1}$	$8 \times 10^{-2}$	0.78
Team 282	$5 \times 10^{-1}$	$6 \times 10^{-1}$	0.24
Team 280	$6 \times 10^{-1}$	$8 \times 10^{-1}$	0.17

\*The *p*-value for this performance was below the precision of our calculation.  
doi:10.1371/journal.pone.0009202.t009

expression of B) and that deletion of direct regulators produces larger changes in expression than deletion of indirect regulators. Then, a network prediction is achieved by taking the absolute value of *z*-score and ranking potential edges from high to low values of this metric. Of note, the *z*-score prediction would have placed second, first, and first (tie) in the 10-node, 50-node, and 100-node subchallenges, respectively.

We do not imply that ranking edges by *z*-score is a superior algorithm for inferring gene regulation networks from null-mutant expression profiles in general, though conditional correlation has its merits. Rather, we interpret the efficacy of *z*-score for reverse-engineering these networks as a strong indication that an algorithm must begin with exploratory data analysis. Because additive Gaussian noise (i.e., simulated measurement noise) is a dominant feature of the data, *z*-score happens to be an efficacious method for discovering causal relationships between gene pairs. Furthermore, *z*-score can loosely be interpreted as a metric for the “information

content” of a node deletion experiment. Subsequently, we will evoke this concept of information content to investigate why some network edges remain undiscovered by the entire community.

**Intrinsic impediments to network inference.** Analysis of the predictions of the community as a whole shed light on two important technical issues. First, are certain edges easy or difficult to predict and why? Second, do certain network features lead teams to predict edges where none exist? We call the former concept the *identifiability* of an edge, and we call the latter concept *systematic false positives*. A straightforward metric for quantifying identifiability and systematic false positives is the number of teams that predict an edge at a specified cutoff in the prediction lists. In the following analysis, we used a cutoff of 2P (i.e., twice the number of actual positives in the gold standard), which means that the first 2P edges were thresholded as present (positives). Incomplete prediction lists were completed with a random ordering of the missing potential edges prior to thresholding.

We grouped the gold standard edges into bins according to the number of teams that identified the edge at the specified threshold (2P). We call the resulting histogram the identifiability distribution (Figure 6A). A community composed of the ten worst-performing teams has an identifiability distribution that is approximately equivalent to that of a community of random prediction lists—the two-sample Kolmogorov-Smirnov test *p*-value is 0.89. By contrast, a community composed of the ten best teams has a markedly different identifiability distribution compared to a random community—the two sample K-S test *p*-value is  $5.3 \times 10^{-27}$ .

The zero column in the identifiability distribution corresponds to the edges that were not identified by any team. We hypothesized that the unidentified edges could be due to a failure of the data to reveal the edge—the problem of insufficient information content of the data. Using the null-mutant *z*-score as a measure of the information content of the data supporting the existence of an edge, we show that unidentified edges tend to have much lower absolute *z*-scores compared to the edges that were identified by at least one team (Figure 6B). This can occur if expression of the target node does not significantly change upon deletion of the regulator. For example, a target node that implements an OR-gate would be expected to have little change in expression upon the deletion of one or another of its regulators. Such a phenomena is more likely to occur for nodes that have a higher in-degree. Indeed, the unidentified edges have both lower *z*-score and higher target node in-degree than the identified edges (Figure 6C).

We investigated whether certain structural features of the gold standard networks led the community to incorrectly predict edges where there should be none. When multiple teams make the same false positive error, we call it a systematic false positive. The number of teams that make the error is a measure of confusion of the community. An ever-present conundrum in network inference is how to discriminate direct regulation from indirect regulation. We hypothesized that two types of topological properties of networks could be inherently confusing, leading to systematic false positives. The first type is what we call shortcut errors, where a false positive shortcuts a linear chain. A second type of direct/indirect confusion is what we call a co-regulation error, where co-regulated genes are incorrectly predicted to regulate one another (see schematic associated with Figure 7).

**Table 10.** Comparison of scores in the 10, 50, and 100-node subchallenges.

<i>In Silico Size 10</i>			<i>In Silico Size 50</i>			<i>In Silico Size 100</i>		
Rank	Team	Score	Rank	Team	Score	Rank	Team	Score
1	<b>B Team</b>	5.12	1	<b>B Team</b>	39.83	1	<b>B Team</b>	Inf*
2	<b>USMtec347</b>	3.82	2	<b>USMtec347</b>	31.34	2	<b>Bonneau</b>	45.44
3	<b>IntigernHSP</b>	2.19	3	Team 256	17.93	3	<b>IntigernHSP</b>	42.24
4	Team 258	2.11	4	<b>Bonneau</b>	15.87	4	Team 305	31.88
5	<b>Bonneau</b>	2.07	5	<b>IntigernHSP</b>	13.77	5	Team 301	11.99
6	Team 305	1.94	6	Team 305	12.20	6	Team 310	8.83
7	Team 256	1.83	7	Team 258	8.40	7	Team 314	8.81
8	Team 110	1.13	8	Team 254	5.12	8	Team 183	8.25
9	Team 273	1.09	9	Team 310	4.84	9	Team 254	7.83
10	Team 183	1.06	10	Team 183	4.13	10	Team 192	4.05
11	Team 310	1.03	11	Team 301	3.56	11	Team 110	2.79
12	Team 282	1.02	12	Team 273	3.42	12	Team 303	2.61
13	Team 86	1.00	13	Team 314	3.15	13	Team 236	2.21
14	Team 303	0.97	14	Team 110	2.32	14	Team 283	2.12
15	Team 70	0.96	15	Team 192	1.90	15	<b>USMtec347</b>	1.99
16	Team 254	0.86	16	Team 303	1.83	16	Team 271	1.75
:	:	:	:	:	:	:	:	:

\*The *p*-value for this performance was below the precision of our calculation.

Best-performers are indicated in bold.

doi:10.1371/journal.pone.0009202.t010

We performed a statistical test to determine if there is a relationship between systematic false positives and the shortcut and co-regulated topologies (Figure 7). Fisher's exact test is a test of association between two types of classifications. First, we classified all negatives (absence of edges) by network topology as either belonging to the class of shortcut and co-regulated node pairs, or not. Second, we classified negatives by the predictions of the community as either systematic false positives, or not. Finally, we constructed the  $2 \times 2$  contingency table, which tabulates the number of negatives classified according to both criteria simultaneously.

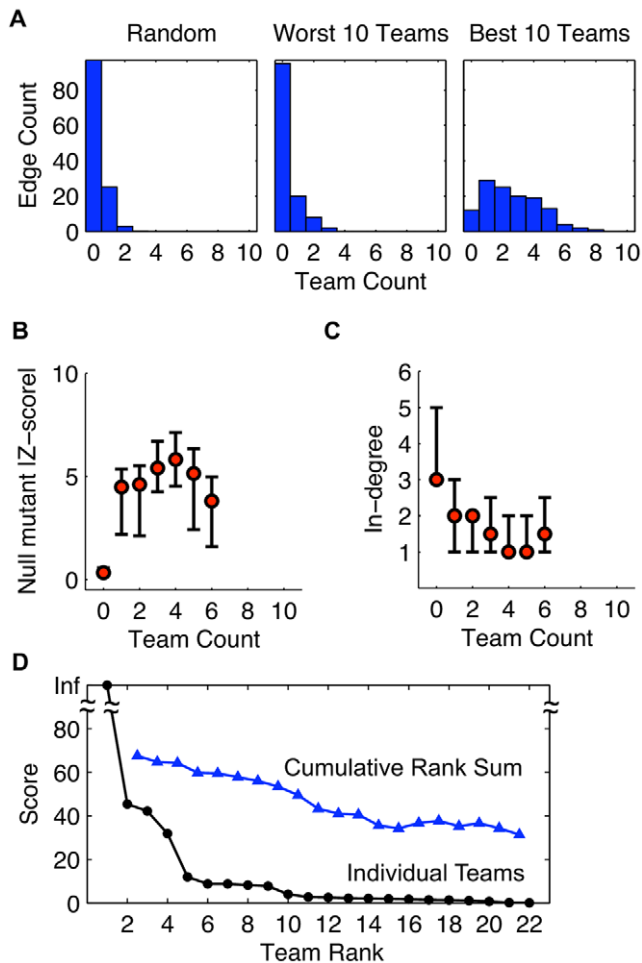
There is a strong relationship between systematic false positives and the special topologies that we investigated. The systematic false positives are concentrated in the shortcut and co-regulated node pairs. This can be seen by inspection of each  $2 \times 2$  contingency table. For example, systematic false positives (the most common false positive errors in the community) have a ratio of 1.09 (51 special topologies to 47 generic topologies) whereas the less common false positive errors have a ratio of 0.11 (920 special topologies to 8757 generic topologies)—a profound difference in the topological distribution of false positives depending on whether many teams or few (including none) made the

error. Direct-indirect confusion of this kind explains about half of the systematic false positives in the Ecol1 network, and more than half in the other 100-node networks.

**Community intelligence.** Does the community possess an intelligence that trumps the efforts of any single team? To test this hypothesis we experimented with various ways of combining the predictions of multiple teams into a consensus prediction. Based on simplicity and performance, we settled on the rank sum. The order of the edges in a prediction list is a ranking. We summed the ranks for each edge given by the various teams, then re-ranked the list to produce the consensus network prediction. Depending on which teams are included, this procedure can boost the overall score. For example, combining the predictions of the second and third-place teams achieved a better score than the second place team (Figure 6D). This result seems to indicate that the predictions of second and third-place teams are complementary; probably these teams took advantage of different features in the data. However, combining predictions with those of the best-performer only degraded the best score. Obviously, if the best prediction is close to optimal, combination with a suboptimal prediction degrades the score.

Starting with the second place team and including progressively more teams, the rank sum prediction score degrades much slower than the score of the individual teams (Figure 6D). This is reassuring since, in general, given the output of a large number of algorithms, we may not know which algorithms have efficacy. The rank sum consensus prediction is robust to the inclusion of random prediction lists (the worst-performing teams predictions were equivalent to random). It seems to be efficacious to blend the results of a variety of algorithms that approach the problem from different perspectives. We expect hybrid strategies to become more common in future DREAM challenges.

**Lessons for experimental validation of inferred networks.** This challenge called for the submission of a ranked list of predicted edges from most confidence to least confidence that an edge is present in the gold standard. Ranked lists are common for reporting the results of high-throughput screens, whether experimental (e.g., differential gene expression, protein-protein interactions, etc.) or computational. In the case of computational predictions, it is typical to experimentally validate a handful of the most reliable predictions. This amounts to characterizing the precision at the top of the prediction list. The outcome of the *in silico* network inference challenge



**Figure 6. Analysis of the community of teams reveals characteristics of identifiable and unidentifiable network edges.** The number of teams that identify an edge at a specified cutoff is a measure of how easy or difficult an edge is to identify. In this analysis we use a cutoff of 2P (i.e., twice the number of actual positive edges in the gold standard network). (a) Histograms indicate the number of teams that correctly identified the edges of the gold standard network called *InSilico\_Size100\_Ecoli1*. The ten worst teams in the 100-node sub-challenge identified about the same number of edges as is expected by chance. By contrast, the ten best teams identified more edges than is expected by chance and this sub-community has a markedly different *identifiability distribution* than random. Still, some edges were not identified by even the ten best teams (see bin corresponding to zero teams). Unidentified edges are characterized by (b) a property of the measurement data and (c) a topological property of the network. (b) Unidentified edges have a lower null-mutant absolute  $z$ -score than those that were identified by at least one of the ten best teams. This metric is a measure of the information content of the measurements. (c) Unidentified edges belong to target nodes with a higher in-degree than edges that were identified by at least one of the ten best teams. Circles denote the median and bars denote upper and lower quartiles. Statistics were not computed for bins containing less than four edges. (d) The benefits of combining the predictions of multiple teams into a consensus prediction are illustrated by the rank sum prediction (triangles). Though no rank sum prediction scored higher than the best-performer, a consensus of the predictions of the second and third place teams boosted the score of the second place team. Rank sum analysis shown for the 100-node sub-challenge.

doi:10.1371/journal.pone.0009202.g006

reveals two reasons why a “top ten” approach to experimental validations is difficult to interpret.

Experimental validations of the handful of top predictions of an algorithm would be useful if precision were a monotonically decreasing function of the depth  $k$  of the prediction list. The actual P-R curves illustrate that this is not the case. In

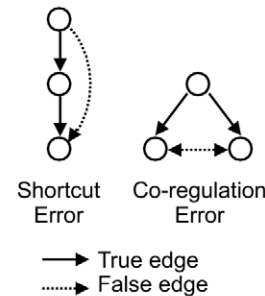
Figure 5A, the best-performer initially had low precision which rose to a high value and was maintained to a great depth in the prediction list. The second-best-performer initially had high precision, which plummeted abruptly with increasing  $k$ . Validation of the top ten predictions would have been overly pessimistic in the former case, and overly optimistic in the

latter case. Unfortunately, since precision is not necessarily a monotonically decreasing function of  $k$ , a small number of experimental validations at the top of the prediction list can not be extrapolated.

**Year-over-year comparison.** We would like to know if predictions are getting more accurate from year to year, and if teams are improving. With only two years of data available, no definitive statement can be made. However, there is one interesting observation from the comparison of individual teams’ year-over-year scores. We compared the results of the 50-node subchallenge of DREAM3 to the results of the 50-node subchallenge of DREAM2 (the subchallenge that was substantially similar from year to year). It is a curious fact that teams that scored high in DREAM2 did not score high in DREAM3. There can be many reasons for the counter-trend. The *in silico* data sets were generated by different people from year to year. Furthermore, the topological characteristics of the networks were different. For example, all of the DREAM3 networks were devoid of cycles whereas the DREAM2 networks contained more than a few. The dynamics were implemented using different, though qualitatively similar equations. Finally, the current year data included additive Gaussian noise, whereas the prior data sets did not. Given the efficacy of directly acknowledging the measurement noise in the reverse engineering algorithm (e.g., null mutant  $z$ -score described above), any team that did not acknowledge the noise would have missed an important aspect of the data. We interpret the year-over-year performance as an indication that no algorithm is “one-size-fits-all.” The *in silico* network challenge data was sufficiently unique from year to year to warrant a custom solution. A final note, teams may have changed their algorithms.

**Survey of methods.** A voluntary survey was conducted at the conclusion of DREAM3 in which 15 teams provided basic information about the class of methods used to solve the challenge (Figure 8). The two most common modeling formalisms were Bayesian and linear/nonlinear dynamical models, which were equally popular (7 teams). Linear regression was the most popular data fitting/inference technique (4 teams); statistical (e.g., correlation) and local optimization (e.g., gradient descent) were the next most popular (2 teams). Teams that scored high tended to enforce additional constraints, such as minimization of the L1 norm (i.e., a sparsity constraint). Also, high-scoring teams did not ignore the null-

		Network Topology		A to B Ratio	P-value	Network
		(A) Shortcut or co-regulated node pairs	(B) All other unconnected node pairs			
Predictions of community	Systematic FP	51	47	1.09	5.1 10 <sup>-26</sup>	Ecoli1
	Rare FP	920	8757	0.11		
	Systematic FP	51	47	1.09	1.4 10 <sup>-20</sup>	Ecoli2
	Rare FP	1223	8460	0.14		
	Systematic FP	41	56	0.73	1.9 10 <sup>-26</sup>	Yeast1
	Rare FP	501	9136	0.05		
	Systematic FP	75	20	3.75	2.4 10 <sup>-22</sup>	Yeast2
	Rare FP	2842	6574	0.43		
	Systematic FP	59	34	1.74	6.5 10 <sup>-8</sup>	Yeast3
	Rare FP	3313	5943	0.56		



**Figure 7. Community analysis of systematic false positives.** Systematic false positive (FP) edges are the top one percent of edges that were predicted by the most teams to exist, yet are actually absent from the gold standard (i.e., negative). Rare false positive edges are the remaining 99 percent of edges that are absent from the gold standard network. The entries of each two-by-two contingency table sum to the total number of negative edges (i.e., those not present) in the gold standard network. There is a relative concentration of FP errors in the shortcut and co-regulated topologies, as evidenced by the A-to-B ratio. *P*-values for each contingency table were computed by Fisher's exact test, which expresses the probability that a random partitioning of the data will result in such a contingency table.  
doi:10.1371/journal.pone.0009202.g007

mutant data set. The main conclusion from the survey of methods is that there does not seem to be a correlation between methods and scores, implying that success is more related to the details of implementation than the choice of general methodology.

## Conclusions

A macro-level goal of the DREAM project is to discover new biological knowledge from the aggregate efforts of the challenge participants. So far, we have not realized this lofty goal although we

believe it will be possible for new knowledge to emerge from future DREAM challenges. This will require that teams build models that are simultaneously predictive and interpretable. Currently, the models offered in response to the DREAM challenges seem to be one or the other, but not both simultaneously. This is reasonable since the DREAM3 challenges solicited either measurement predictions or network predictions, but not both. Some of the DREAM4 challenges, which as of this writing are underway, attempt to remedy this disconnect.

Predicting measurements falls within the classic statistical learning paradigm whereby a training set is used to learn a model and a test set is used to evaluate how well the model generalizes. Regression-type methods performed well in this type of challenge. By comparison, biological network inference is less of a proven science. Perhaps it is this “wild west” character that attracts such high levels of participation in the DREAM challenges. The best-performer in the *in silico* network inference challenge appropriately handled measurement noise after exploring the

	10-node	50-node	100-node	WT	Null	Hetero	Trajectory	Boolean	Bayesian	Linear Dyn	Nonlin Dyn	Other	Statistical	Info Theory	Lin Regress	Local Opt	Global Opt	Other	In-degree	Out-degree	Other
B Team	5.124	39.83	Inf	X	X	X	X			X	X	X	X			X					
Bonneau	2.07	15.87	45.44	X	X	X	X		X					X							L1 sparsity
IntigernHSP	2.19	13.77	42.24	X	X	X	X				X						X				
USMtec347	3.82	31.34	1.99	X	X	X	X	X			X						DBN				X
Team 301	0.80	3.56	11.99	X	X	X	X			X				X							
Team 310	1.03	4.84	8.83	X	X	X		X					X			X			5		
Team 256	1.83	3.15	8.81	X	X	X	X			X		X			X				2		
Team 254	0.86	5.12	7.83	X	X	X			X								Lasso				L1 norm
Team 192	0.61	1.90	4.05	X		X	X	X	X												
Team 236	0.84	1.32	2.21				X	X				X							X		
Team 283	0.24	1.67	2.12	X	X	X	X	X	X												
Team 70	0.96	1.25	1.41			X			X					X					X		
Team 302	0.39	0.65	1.17	X		X			X					X							
Team 280	0.29	0.35	0.17		X			X													
Team 190	0.83	0.57					X	X		X					X						
	Score			Data				Model Type				Inference Method						Constraints			

**Figure 8. Survey of *in silico* network methods.** There does not seem to be a correlation between methods and scores, implying that success is more related to the details of implementation than the choice of general methodology.  
doi:10.1371/journal.pone.0009202.g008

character of the data. *Ad hoc* procedures based on exploratory data analysis seem to be rewarded by the *in silico* network inference challenge.

### Lessons Learned from Network Inference

After pouring over the predictions of the systems biology modeling community we have learned one overriding lesson about modeling and prediction of intracellular networks. There is no such thing as a one-size-fits-all algorithm. An algorithm has no intrinsic value in isolation of the data that motivated its creation. DREAM identifies the best teams with respect to specific challenges, not the best algorithms. This is an important distinction to keep in mind when interpreting results, especially results of the *in silico* challenge where the data is admittedly non-biological despite our best efforts. The matching of algorithm to data is fundamental for efficacy.

It would be inappropriate to dismiss an algorithm on the basis of a lackluster DREAM score. As a sanity check, we ran a well-respected network inference algorithm on the *in silico* data set. We do not name the algorithm or its authors, in keeping with one of the founding principles of DREAM—do no harm. Surprisingly, the algorithm, which is described or applied in string of high-profile publications, did not make statistically significant network predictions. Upon further examination of the data, we realized that the signal required by this particular algorithm was nearly absent from the *in silico* data set.

The perturbations used in the *in silico* data set are inappropriate for methods that expect pairs of linked nodes to covary under many conditions (e.g., correlation-based methods). In this data set, parent node mutations resulted in large expression changes in their direct targets. However, small expression changes due to indirect effects were not prominent. This is why null-mutant *z*-score was an efficacious signal for network inference, but measures of statistical dependence were not.

The Boolean-like character of the data was probably a consequence of the fact that there were no feedback loops in any of the fifteen *in silico* networks, which were extracted as subgraphs of the known *E. coli* and *S. cerevisiae* gene regulation networks. Although it is true that explicit feedbacks are exceedingly rare in the transcriptional networks of these organisms, there is extensive feedback control exerted at the level of protein-protein and protein-metabolite interactions, which was not represented in the generative model. Neverthe-

less, the generative model used in this challenge may provide a qualitatively accurate depiction of unicellular gene expression.

Since parent-child correlations are present in certain real-world gene expression data sets where correlation-based algorithms have been successfully applied, we cannot conclude that measures of statistical dependence are a poor choice for reverse-engineering gene networks in general, only for this particular data set. The *in silico* challenge seems to reward strategies that are guided by exploratory analysis of the data itself and that are adapted to the perturbations that exercise the network, and presumably penalizes established algorithms which may be based on different characteristics of certain real data sets.

Another take-home lesson from DREAM3 is that top-ten style validations are nearly impossible to interpret because precision is rarely a monotonically decreasing function of the threshold *k*. The best-performer and runner-up in the 100-node subchallenge illustrate the problem (see P-R curve in Figure 5B). The best-performer was identified because the overall network reconstruction was highly significant despite that the “high confidence” edge predictions were actually quite poor. By contrast, the runner-up in the 100-node subchallenge had very high precision for the high-confidence edge predictions but the overall network reconstruction was many orders of magnitude less significant than that of the best-performer. If the goal is to make a handful of predictions which are to be investigated by follow-up experiments, the latter situation is clearly desirable. If the goal is to sketch a large-scale network, the former situation is best. Both results are potentially desirable depending on the question at hand. However, in either case, validation of the top predictions gives no information about the overall trajectory of precision deeper into the edge list, and thus no information about the overall quality of the reconstructed network. Related to this last point, it may be that the best course of action is to combine the predictions of algorithms that focus on different aspects of the data into a consensus prediction, for example, by summing ranks and then re-ranking as we have illustrated in Figures 4B and 6D.

### Lessons Learned from Prediction of Measurements

The signaling response and gene expression prediction challenges were presented in typical machine learning style

(e.g., training set/prediction set). Teams that adopted and embellished standard methods in machine learning such as the lasso (a variant of subspace regression) and *k*-nearest neighbors (a type of local estimation) predicted protein and mRNA expression well. It is interesting to ponder if prediction of measurements (from other measurements) may be relevant to experimental design of high-throughput molecular profiling studies.

We desire that models are interpretable in terms of plausible biology. However, extensive knowledge of signal transduction pathways were intentionally ignored by the best-performers in the signaling response prediction challenge. This turned-out to be a winning strategy, speaking to the power of statistical methods for prediction. Likewise, one of the best-performers in the gene expression prediction challenge ignored the known *S. cerevisiae* gene regulation network, but the other best-performer took advantage of it.

As evaluators, we do not have access to the models, only the predictions. Unfortunately, the predictions do not seem to deliver on our overarching goal of learning about the biology of signal transduction or gene expression. Even if we had access to the models, there is a dearth of interpretable biology to be learned from nearest-neighbor and regression approaches. In future runs of DREAM, we will strive to enhance the interpretability of the community's prediction efforts by incentivizing interpretability of models in terms of plausible biology. We believe that this community will rise to the challenge. It is interesting to ponder whether biologically plausible/interpretable models are at a disadvantage compared to regression-like methods when it comes to predicting measurements.

### Caution and Optimism

The vast majority of the teams' predictions were statistically equivalent to random guesses. It is likely that some of these teams employed methods that have previously been published and experimentally validated. This does not trouble us, since every algorithm is born from exploratory data analysis, so off-the-shelf applications of published algorithms would not be expected to perform well on fresh data sets. This has implications for the marketing of algorithms as being fit for a specific purpose, like gene regulation network inference. Even for this particular problem there is no one-size-fits-all algorithm. If algorithms *are* to be applied off-the-shelf, a strategy for guarding against bogus predictions is to employ a wide variety of



algorithms that train on different features of the data (e.g., correlation [28], mutual information [29–32], synergy [33], etc.), then merge the results to gain confidence in the predictions.

The ill-posed signaling cascade identification challenge drew some fair criticism from participants who felt misled. One participant commented,

For me, the take-home message is that if you want to build a mathematical model to explain a dataset, you should have a good understanding of the dataset. In other words, modelers and experimentalists need to collaborate closely. I think that's the main problematic difference between DREAM and CASP. For [protein] structure determination

the experimental and theory sides are relatively separable, but for network analysis the set of questions you can ask is so broad that the two sides need to work together to figure out both what experiments to do and how to analyze them afterwards.

The failure in execution of this challenge was due to a communication breakdown between the experimentalist who provided the data and the DREAM organizers. Had we, the organizers, been more experienced in the technical details of immunolabeling, the challenge would have written unambiguously. To the data producer, there was no ambiguity.

As challenge designers, we desire models that are simultaneously predictive and

interpretable. Future runs of DREAM will encourage simultaneous submissions of networks and predictions, which may help us close-in on the macro-level goal of DREAM, to learn new biology from the aggregate predictions of the systems biology modeling community.

The exemplary predictions of the best-performers in each challenge are cause for celebration. As mentioned above, DREAM identifies the best teams, not the best algorithms. The importance of exploratory data analysis can not be stressed enough. The best-performers are exceedingly talented at their craft.

## Acknowledgments

We thank all participants of the challenge for their contribution. We thank Andrea Califano and Mukesh Bansal for valuable discussions.

## References

- Kim HD, Shay T, O'Shea EK, Regev A (2009) Transcriptional regulatory circuits: predicting numbers from alphabets. *Science* 325: 429–432.
- Kremling A, Fischer S, Gadkar K, Doyle FJ, Sauter T, et al. (2004) A benchmark for methods in reverse engineering and model discrimination: problem formulation and solutions. *Genome Res* 14: 1773–1785.
- David LA, Wiggins CH (2007) Benchmarking of dynamic bayesian networks inferred from stochastic time-series data. *Ann N Y Acad Sci* 1115: 90–101.
- Camacho D, Vera Licona P, Mendes P, Laubenbacher R (2007) Comparison of reverse-engineering methods using an in silico network. *Ann N Y Acad Sci* 1115: 73–89.
- Cantone I, Marucci L, Iorio F, Ricci MA, Belcastro V, et al. (2009) A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell* 137: 172–181.
- Haynes BC, Brent MR (2009) Benchmarking regulatory network reconstruction with *grendel*. *Bioinformatics* 25: 801–807.
- Hache H, Wierling C, Lehrach H, Herwig R (2009) Genge: systematic generation of gene regulatory networks. *Bioinformatics* 25: 1205–1207.
- Marbach D, Schaffter T, Mattiussi C, Floreano D (2009) Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *J Comput Biol* 16: 229–239.
- Stolovitzky G, Monroe D, Califano A (2007) Dialogue on reverse-engineering assessment and methods: the dream of high-throughput pathway inference. *Ann N Y Acad Sci* 1115: 1–22.
- Stolovitzky G, Prill RJ, Califano A (2009) Lessons from the DREAM2 challenges. *Ann N Y Acad Sci* 1158: 159–195.
- Moult J (2006) Rigorous performance evaluation in protein structure modelling and implications for computational biology. *Philos Trans R Soc Lond B Biol Sci* 361: 453–458.
- (2009) DREAM website. <http://wiki.c2b2.columbia.edu/dream/>.
- Feinerman O, Veiga J, Dorfman JR, Germain RN, Altan-Bonnet G (2008) Variability and robustness in t cell activation from regulated heterogeneity in protein levels. *Science* 321: 1081–1084.
- Saez-Rodriguez J, Alexopoulos LG, Epperlein J, Samaga R, Lauffenburger DA, et al. (2009) Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Mol Syst Biol* 5: 331.
- Alexopoulos L, Saez-Rodriguez J, Cosgrove B, Lauffenburger D, PK S (2009) Networks reconstructed from cell response data reveal profound differences in signaling by tolllike receptors and nf- $\kappa$ b in normal and transformed human hepatocytes. Under review.
- Saez-Rodriguez J, Goldsipe A, Muhlich J, Alexopoulos LG, Millard B, et al. (2008) Flexible informatics for linking experimental data to mathematical models via datarail. *Bioinformatics* 24: 840–847.
- Natarajan K, Meyer MR, Jackson BM, Slade D, Roberts C, et al. (2001) Transcriptional profiling shows that *gcn4p* is a master regulator of gene expression during amino acid starvation in yeast. *Mol Cell Biol* 21: 4347–4368.
- Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4: 249–264.
- Granek JA, Clarke ND (2005) Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biol* 6: R87.
- Guex N, Migliauacca E, Xenarios I (in press) Multiple imputations applied to the dream3 phosphoproteomics challenge: A winning strategy. *PLoS ONE*.
- Bourque G, Clarke N (submitted) Success in the dream3 signaling response challenge using simple weighted-average imputation: lessons for community-wide experiments in systems biology. *PLoS ONE*.
- Gustafsson M, Hornquist M (in press) Gene expression prediction by soft integration and the elastic net. *PLoS ONE*.
- Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B* 67: 301–320.
- Jianhua R (submitted) A top-performing algorithm for the dream3 gene expression prediction challenge. *PLoS Biol*.
- Yip K, Alexander R, Yan K, Gerstein M (in press) Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data. *PLoS ONE*.
- Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, et al. (2006) The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol* 7: R36.
- Madar A, Greenfield A, Vanden-Eijnden E, Bonneau R (submitted) Network inference using dynamic context likelihood of relatedness and inferelator 1.0. *PLoS ONE*.
- Rice JJ, Tu Y, Stolovitzky G (2005) Reconstructing biological networks using conditional correlation analysis. *Bioinformatics* 21: 765–773.
- Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS (2000) Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *Proc Natl Acad Sci U S A* 97: 12182–12186.
- Butte AJ, Kohane IS (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*. pp 418–429.
- Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, et al. (2005) Reverse engineering of regulatory networks in human b cells. *Nat Genet* 37: 382–390.
- Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, et al. (2006) Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7 Suppl 1: S7.
- Anastassiou D (2007) Computational analysis of the synergy among multiple interacting genes. *Mol Syst Biol* 3: 83.